Structural Combination of Seasonal Exponential Smoothing Forecasts Applied to Load Forecasting

Juan F. Rendon-Sanchez

Cass Business School, City, University of London, 106 Bunhill Row, London EC1Y 8TZ, UK

Lilian M. de Menezes*

Cass Business School, City, University of London, 106 Bunhill Row, London EC1Y 8TZ, UK

Abstract

This article draws from research on ensembles in computational intelligence to propose structural combinations of forecasts, which are point forecast combinations that are based on information from the parameters of the individual models that generated the forecasts. Two types of structural combination are proposed which use seasonal exponential smoothing as base models, and are applied to forecast short-term electricity demand. Although forecasting performance may depend on how ensembles are generated, results show that the proposed combinations can outperform competitive benchmarks. The methods can be used to forecast other seasonal data and be extended to different types of forecasting models.

Keywords: Forecasting, combination of forecasts, electricity demand/load forecasting, ensembles, exponential smoothing methods

*Corresponding author

Email addresses: juan.rendon.1@cass.city.ac.uk (Juan F. Rendon-Sanchez), L.deMenezes@city.ac.uk (Lilian M. de Menezes)

1. Introduction

It is now several decades since Bates & Granger (1969) demonstrated that a combination of forecasts can outperform individual forecasts. Since then several methods have been developed, and reviews of the literature (Clemen, 1989; De Menezes et al., 2000) concluded that combining multiple forecasts tends to increase forecast accuracy. As Timmermann (2006) argued, unless there is a forecasting model that is consistently more accurate than its competitors, a combination of forecasts enables diversification and is thus better than relying on a single forecast for decision making. There are different approaches to forecast combination, one of which is the creation of ensembles of forecasting models whose forecasts are then averaged, and is a source of inspiration for this research.

1.1. Traditional combinations of forecasts and extensions

Stock & Watson (2004) have observed that often the best performing combinations are simple, and averaging point forecasts has become a practice advocated in textbooks (Ord et al., 2017) and forecasting principles (Armstrong, 2001). Linear combination is one of the simplest approaches, yet the simple average is difficult to defeat (Armstrong, 2001). Che (2015) suggested improvements to the selection of models for linear combinations by using the concept of entropy to minimise linear redundancy and maximise linear relevance. Chan et al. (2004) examined whether and when the weights in a combination should change, and concluded that there are benefits from time-varying weights and complex models. While focusing on forecast error distributions and their dynamics, De Menezes et al. (2000) observed that gains could be achieved from averaging large number of forecasts (De Menezes & Bunn, 1998), and specially in cases where forecast errors were serially correlated, the simple average was advocated.

Clustering forecasts, rather than averaging, is inspired by the assumption of commonalities underlying forecasting models. K-means has been proposed, for example, as an approach to group similar forecasting models into clusters based on past Mean Square Error (MSE) of individual forecasts (see Timmermann, 2006). Switching between different forecasts at different periods (Granger, 1993; Deutsch et al., 1994; Taylor & Majithia, 2000) is a selection strategy that makes use of information from different forecasts. It assumes that available forecasts might vary in relevance depending on the period to forecast. Kolassa (2011) combined exponential smoothing forecasts and incorporated information criteria when calculating weights.

1.2. Ensembles

The rationale for using ensembles of univariate models, is that by slightly varying initial conditions and parameters in a model, different predictions can be obtained and a summary of their distribution, such as their mean or median, can be used as a forecast. This method involves three main steps: generation, pruning and combination.

Ensembles of neural networks (NN) are widely used, in particular to forecast short-term electricity demand, which is a problem where the effort of generating many different models is judged worthwhile (e.g. Barak & Sadegh, 2016; Li et al., 2016; Khwaja et al., 2017). The approach has also been extended to other methods. Combinations using ensemble forecasts provided by climate agencies have been proposed, for example, by Taylor & Buizza (2003) and the building of ensembles containing statistical models (ARMA) has been adopted in several applications, e.g. Matijaš et al. (2013). Variations in data patterns imply the use of different forecasting algorithms. Hence, a wider set of models of different complexity have been considered in ensembles, including statistical ARIMA and artificial intelligence methods (Multilayer Perceptron and Recurrent Neural Networks, and support vector machines). The literature underscores the usefulness of integrating computational intelligence and traditional forecasting approaches.

Kourentzes et al. (2014) constructed ensembles of exponential smoothing forecasts by extracting components from the original time series through temporal aggregation, fitting exponential smoothing models to these and combining their forecasts. Their empirical evaluation demonstrated significant improvements in forecasting accuracy, especially for long-term forecasts. More recently, Bergmeir et al. (2016) created different replicas of a time series by using a Box-Cox transformation followed by a sophisticated time-series decomposition, such that the remainder of the series was bootstrapped. Ensembles of exponential smoothing models were fitted leading to a forecast. Through this procedure, the original exponential smoothing models were outperformed consistently. The exploration of these encouraging results has been carried forward by Petropoulos et al. (2018), who assessed the conditions of success of the approach.

1.3. Extensions of ensembles, hybrid models, and the use of model structure in combinations

Hybrid foresting approaches have been proposed, where statistical and computational intelligence methods are integrated. Genetic algorithms (GA) have been used, in this context, as part of either the optimisation process or the model specification mechanism (see for example Zhou et al., 2002; Pai & Hong, 2005). Fuzzy inference systems were also investigated while combining forecasts (Fiordaliso, 1998; Xiong et al., 2001), because their ability to find non-linear mappings between an input and an output space can be a combining mechanism. Non-linear combinations of forecasts have tended to use NNs, because of perceived advantages over linear combination schemes (Donaldson & Kamstra, 1996). In general, the importance of combinations of computational intelligence models has been highlighted by Crone et al. (2011) in the context of the NN3 forecast competition. Several of the most competitive models in NN3, contained a form of combination of forecasts.

Structural information of a model can relate to its functional form or to the specific values of its parameters. Such information has now been used by different studies in forecast combinations, although some authors have not explicitly acknowledged their use. Indeed, Bakker & Heskes (2003), to the best of our knowledge were the first to argue that, in ensembles of NNs, a combination could be achieved by exploring diversity not only in individual forecasts, but also in the parameter space of the models that led to the individual forecasts. Their study explored diversity via clustering for model selection, thus taking the view that not all forecasts in an ensemble should be included in the combination. Another possibility to use the structure of NNs is to explore relationships between components of the model and its forecasts. For example, in ensembles of NNs, Garson (1991) and Goh (1995) estimated the importance of an input variable by examining the weighted connections between nodes of interest from the input to the output, and constructing in this way mappings of importance, which can be used to estimate weights for forecast combinations.

Kolassa (2011), when using Akaike weights in their combinations, incorporated parameter information indirectly, since the combining weights are based on the number of parameters in the model. Bergmeir et al. (2016) fitted different bootstrapped time series to ETS state-space models (Hyndman & Athanasopoulos, 2014). In doing so, they automatically created model parameter diversity and the forecasts combinations are, therefore, achieved by taking into account parameter information in the process. This is similar to the procedure adopted by Kourentzes et al. (2014). In their work, however, both parameter values and functional form were taken into account, since the models selected by ETS routines can vary.

Bakker & Heskes (2003) considered the possibility of clustering NNs in their parameter space. The present study takes inspiration from them in order to combine forecasts of single-seasonal and double-seasonal exponential smoothing. In terms of structural information from the models, only parameter values are here considered. The single seasonal Holt-Winters considered here belongs to the family of exponential smoothing methods investigated by Bergmeir et al. (2016). The focus on double-seasonal time series is particularly important for short-term electric load forecasting. Additionally, the structural (parameter) information is here explicitly incorporated in the modelling procedure, by performing either clustering or genetic algorithms searches in the parameter space in order to select forecasts to be combined.

1.4. Load forecasting and combining forecasts

The safe and efficient operation of power systems requires accurate forecasts of electricity demand and of the expected load in the electric system. In electricity markets, forecasts of the demand for electricity are also critical to support transactions and decision making, because the limited ability to store electricity implies that prices are very volatile, especially as the interval between the transaction and the delivery of power decreases. Forecasting methods that have been used are broadly categorised as: classical time series and regressions, artificial and computational intelligence methods, and hybrid approaches (Hahn et al., 2009). The latter includes different types of combinations which, however, tend to use models within a single category. Nowotarski et al. (2016) has combined models by varying training configurations, such as data partitioning, and the choice of the threshold to stop the estimation process. Models with overlapping configurations are averaged in different ways, including OLS and machine learning techniques. These variation induction mechanisms can be traced to Pesaran & Pick (2011), who combined forecasts produced by changing the estimation windows. As a whole, methods to forecast electricity demand in the short term aim to capture the seasonality and/or factors that drive consumption of electricity. Hippert et al. (2001), while reviewing this literature, highlighted the relative success of complex neural networks in addressing the problem. Yet, the findings of (Clements & Harvey, 2010) and those by Taylor et al. (2006), whose comparison included the first two types of methods, seem to favour classical time series and regression models. In fact, Hahn et al. (2009) concluded that future research in this area should focus on hybrid models and explore the potential from different approaches to forecasting.

In this study, three applications with varying complexity illustrate the proposed methodology. The first considers the problem of forecasting peak hourly electricity demand and explores combinations of multiplicative Holt-Winters models, which is commonly used to forecast seasonal time series and is available in most statistical software. The second and third address the problems of forecasting hourly and half-hourly electricity demand, respectively, as intra-day and weekly seasonality patterns are present in the data, the double-seasonal Holt-Winters-Taylor model (Taylor, 2003) is considered.

The remainder of this article is organised as follows. Section 2 describes the base models and how they were prepared for the ensembles, the combination approach, and how the applications were set. Section 4 reports the results. Section 5 concludes and suggests directions for future research.

2. Methodology

Figure 1 summarises the proposed approach to combine forecasts. Accordingly, replicas of the time series data are generated through a variation induction mechanism (block swapping or noise addition) and the base models are fitted to them, thus creating a pool of diverse exponential smoothing models. This pool ($HW_1 \ldots HW_n$) is subject to what we will call a structural combination procedure and forecasts are produced.

2.1. Generating replicas of the time series

In order to structurally combine forecasts, model diversity is required, and thus, prior to fitting the exponential smoothing models to the data,

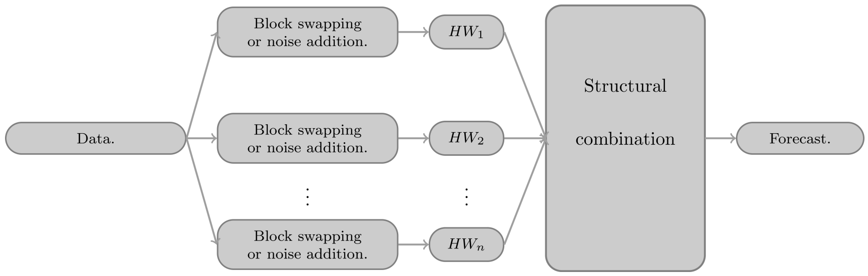

Figure 1: Method

replicas of the original time series (Data) are generated. Two variation induction mechanisms that are commonly used in ensembles are adopted: noise addition (Brown et al., 2003; Zhang, 2007) and block bootstrapping (Jing, 1997). The rationale for noise addition is that any given time series can be regarded as a single realisation from an infinite set of possible realisations, which can be captured by adding a normally distributed noise with mean zero and a relatively small standard deviation.

Zhang (2007) argued that resampling procedures which maintain the serial correlation in the time series can also be adopted. Hence, b blocks of data are defined, i.e.: $z_t = (y_t, \ldots, y_{t+k-1})$ of length k from the original time series $(y_1, y_2, \ldots, y_T)$, where $b = T - k + 1$; and by sampling with replacement from blocks $(z_1, z_2, \ldots, z_b)$ other samples $(z_1^*, z_2^*, \ldots, z_l^*)$ are produced (Zhang, 2007, p. 5333). However, rather than building entire series from blocks of data taken from the original series, randomly selected pairs of data blocks are here swapped. The block size is equal to the longest seasonal cycle and the number swaps is kept low, thus preserving the structure of the series. This means that we can assume that the replicas and the original series are samples from the same data generating process.

2.2. The base models

The first base model considered is derived from standard Holt-Winters method (Hyndman et al., 2008) with an added autoregressive error correction term. This model is suitable for seasonal time series and is also supported by previous studies on forecasting electricity demand (e.g. Clements & Harvey, 2010; Taylor et al., 2006). Forecasts are produced up to S_1 steps ahead, according to the following:

$$\begin{aligned} l_t &= \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}) \\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \\ s_t &= \gamma \frac{y_t}{l_{t-1} + b_{t-1}} + (1-\gamma)s_{t-m} \\ \hat{y}_{t+h|t} &= (l_t + b_t h)s_{t-m+h} + \phi^h(y_t - (l_{t-1} + b_{t-1})s_{t-m}) \end{aligned} \tag{1}$$

where α, β and γ are smoothing parameters; s_t is the seasonal index, b_t represents the trend, l_t the level; m is the season length, $\hat{y}_{t+h|t}$ is the h step-ahead forecast from forecast origin t and ϕ is the parameter of the autocorrelation error correction.

The second base model follows from the multiplicative Holt-Winters-Taylor exponential smoothing method by Taylor (2003), which captures the double seasonality in electricity demand time series of a higher frequency:

$$\begin{aligned} l_t &= \alpha \left(\frac{y_t}{S1_{t-m_1} S2_{t-m_2}} \right) + (1-\alpha)(l_{t-1} + b_{t-1}) \\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \\ S1_t &= \gamma \frac{y_t}{l_t S2_{t-m_2}} + (1-\gamma)S1_{t-m_1} \\ S2_t &= \omega \frac{y_t}{l_t S1_{t-m_1}} + (1-\omega)S2_{t-m_2} \\ \hat{y}_{t+h|t} &= (l_t + hb_t)S1_{t-m_1+h}S2_{t-m_2+h} + \phi^h \left(y_t - (l_{t-1} + b_{t-1})S1_{t-m_1}S2_{t-m_2}\right) \end{aligned} \tag{2}$$

l_t and b_t, are the smoothed level and trend. $S1_t$ and $S2_t$ are the seasonal indices for the intra-day and intra-week seasonal cycles, respectively; m_1 and m_2 are the intra-day and intra-week season lengths, respectively; α, β, γ, ω are the smoothing parameters; $\hat{y}_{t+h|t}$ is the h step-ahead forecast made from forecast origin t and ϕ is the parameter of the autocorrelation error correction. Forecasts are produced up to m_1 steps ahead.

The models were implemented in Matlab® 2010. The level, trend and seasonal components were initialised by averaging the early observations through moving average filters. Although Taylor (2010) optimised parameters based on the sum of squared errors (SSE), more recent studies, (e.g. Arora, 2013), minimised the in-sample root mean squared errors (RMSE). The latter approach is adopted in this study.

2.3. Structural combinations from exponential smoothing

The structure of an exponential smoothing model can be represented as a vector containing its parameters (e.g. α, β, γ and ϕ in Equation 1). Each model can be seen as a point in the N-dimensional space, where N is the number of parameters. Suppose that $c(\cdot)$ is a combination of forecasts; f_i is the forecast produced by model $M_i(A_i)$, where M_i is the model and $A_i = [\alpha_i, \beta_i, \gamma_i, \phi_i]$ is a structural descriptor of M_i (parameter values, for example). Then $f_i = M_i(x; A_i)$ where x is an input. One approach is combining based on f_i, that is $c(\{f_i\})$, and a different approach is to combine as $c(\{f_i\}; \{A_i\})$, that is, a combination where parameter information, A_i, is considered.

2.3.1. Combination based on clustering (CB)

Fuzzy C-Means is an algorithm that partitions a collection of vectors into c fuzzy groups and finds a cluster centre in each group so that a cost function of dissimilarity is minimised (Jang et al., 1997). When clustering forecasting models in their parameter space, the output of the algorithm (centroids in the model space) can be used as reference points to select models in their vicinity. The forecasts of selected models can then be combined. We prefer C-means to K-means (used by Bakker & Heskes, 2003, with NNs in a similar approach) because the identification of completely distinct clusters, as implied by K-means, is likely to be unrealistic since each forecasting model attempts to capture the same underlying time series process. An important feature of C-means is the use of a degree of membership to clusters (between 0 and 1), instead of a binary membership (0 or 1, equivalent to *non-member* or *member*). Therefore, a given object (model in this case) may belong to several groups with different degrees of belongingness between 0 and 1. Normalisation can be imposed, such that the summation of degrees of belongingness of an object to all the clusters is always equals unity (as described by Jang et al., 1997, p. 426). However, given that C-means produces non-deterministic clusters, a variant of the algorithm is here adopted (based on Friedman, 1991), which uses a recursive partitioning of objects in their space that enables a deterministic partition (the same partition is obtained when faced with the same set of objects). The centre of those partitions become the cluster centres to be considered in the combinations, as candidate models were selected in their vicinity.

Clustering is performed over pools of forecasting models. When creating pools, for each constructed series (through noise addition or block swapping), the in-sample errors are calculated on the constructed series itself. When the clustering is performed, the selected models in the clusters have their fit evaluated on the original series before being combined.

Given a forecast origin t, each model in the clusters produces forecasts $\hat{y}_{t+1}, \ldots, \hat{y}_{t+h}$. The combined forecast for $t + h$ is calculated based on a combination of the forecasts for this horizon produced by the selected models in each of the n clusters:

$$\hat{y}_{t+h} = \sum_{i=1}^{n} \phi_i \hat{y}_{C_i,t+h} \tag{3}$$

where ϕ_k is an average of the normalised membership degree of models selected within cluster k:

$$\phi_k = \frac{\sum_{m \in C_k} w_m(v_{C_k})}{N_k} \qquad w_i(v) = \frac{u_i(v)}{\sum_{j=1}^{n} u_j(v)} \qquad u_i(v) = e^{-\frac{D_i^2(v)}{\sum_{j=1}^{n} D_j^2(v)}} \tag{4}$$

$w_i(v)$ is the normalised membership degree of v to cluster i and $u_i(v)$ is the membership of v to cluster i (v is a model represented as a vector with its parameters). The squared euclidean distance $D_i^2(v)$ between v and the i-th centre is divided by the sum of squared distances from v to all centres. An exponential transformation is taken, in order to allow the membership of a model in a cluster to decrease as the distance from the centre increases. C_k denotes cluster k, v_{C_k} is the centre of such cluster and N_k is the number of models in it. In Equation 3, $\hat{y}_{C_i,t+h}$ is the output from cluster i for step $t+h$, that is:

$$\hat{y}_{C_i,t+h} = \alpha_{i,0} + \alpha_{i,1}\hat{y}_{i,1,t+h} + \alpha_{i,2}\hat{y}_{i,2,t+h} + \ldots + \alpha_{i,L}\hat{y}_{i,L,t+h} \tag{5}$$

Variables $\hat{y}_{i,1,t+h}, \hat{y}_{i,2,t+h}, \ldots$ represent the forecasts for $t+h$ produced by models selected within cluster i, and $\alpha_{i,j}$ are the coefficients for forecasts obtained with model j from cluster i, including a constant $\alpha_{i,0}$. L models are selected, with L varying between 1 and 5 (refer to section 3 for details on this choice).

The partitioning algorithm, in a forward (growing) step, exhaustively searches each dimension of the model space and tries partitioning it while fixing the others. When a partition results in a reduction of the loss function (RMSE), it is stored, and the search continues. In a backwards (pruning) step, it revisits iteratively the partitions and eliminates one at a time, if the resulting configuration further reduces loss function. During this process, the loss function is calculated as follows: first an OLS regression is estimated, using in-sample data, of the actual values of the series on the forecasts produced by models selected in each cluster, according to Equations 5. In this way, the α coefficients are obtained. Then the ϕ parameters, according to Equations 4, are calculated and forecasts produced. When the growing and pruning processes come to an end, a further optimisation of α and ϕ parameters is performed through a non-linear optimisation routine (*fmincon*[1] in Matlab).

[1]The configuration was: x, the vector to be optimised, comprised $\alpha_{i,j}$. Constraints

2.3.2. Structural combination based on genetic algorithms (GA)

A structural combination based on genetic algorithms (GA) is the second approach proposed. It is a simpler form of combination (average) still informed by the parameter space.

A series of reference points in the model parameter space is generated, which are analogous to cluster centres. From each point, P_i, five models are selected, as those having the smallest euclidean distance to it. The forecasts from these models are averaged, thus producing forecasts for each reference point. The final forecast combination for horizon h ($\hat{y}_{Avg,t+h}$) is the average of the points forecasts produced in the previous step from the reference points. Genetic algorithms are used to select the reference points such that one-step ahead mean squared error (MSE) of the $\hat{y}_{Avg}$ is minimised. Individuals to evolve are then vectors, each of which represents a set of centres in the model parameter space.

The algorithm is run over the same pool that is used in the cluster-based structural combination, and was implemented in Matlab® 2010 using *ga* routine, with a maximum number of generations equal to 3000.

3. Analysis procedure

Table 1 summarises the main design factors of the study. For each time series $N_{Dv} \times N_{Lv} = 6$ pools are created, each one concentrating in a method and a level for data variation. Each pool contains $N_{Mp} = 50$ models, each fitted to a different series replica. In each application, the time series replicas are divided into training and testing periods. The training period is used to fit the models and the testing period to assess the out-of-sample

---

$A \cdot x \leq b$ and $Aeq \cdot x = beq$ were left empty for unconstrained searches; lower and upper bounds for estimates were left open. The maximum number of function evaluations allowed was MaxFunEvals = 3000. The maximum number of iterations allowed was MaxIter = 3000 and tolerance for the objective function is TolFun = 0.0001.

performance.

For each pool, 3 CB and 3 GA combinations are fitted and an average forecast calculated. Therefore, for each series, the number of forecast combinations is $N_{Dv} \times N_{Lv} \times N_C \times N_{Sa} + N_{Dv} \times N_{Lv} = 42$. The combination models are denoted as $CB(M, L_v, N_C)$ or $GA(M, L_v, N_C)$, where M is the data variation method (noise addition or block swapping), L_v is the level of variation and N_C corresponds to how many options are considered in the analysis for the maximum number of clusters: 2, 4 or 8, thus 3 options.

Table 1: Design factors

Factor	Values
N_s: Number of time series	3
N_{Dv}: Number of data variation mechanisms	2
N_{Lv}: Number of levels of data variation	3
N_C: Number of alternative maximum number of clusters	3
N_{Sa}: Number of structural approaches	2 (CB and GA)
N_{Mp}: Number of models in each pool	50
N_{Mc}: Maximum number of models in each cluster	5

N_{Mp} is equal to 50, in accordance with Bakker & Heskes (2003). Having a maximum of 5 models selected per cluster and 2, 4 and 8 clusters allows to use at least 20% of the total and at most 80% of the individual models in the combinations, thus following the advice from previous studies that considered a fraction of models in the pool to be included in the ensemble (e.g. Zhou et al., 2002). This reasoning was applied to both the clustering-based and the genetic-algorithm-based approaches.

Three levels of block swapping are used: $0.1I$, $0.2I$ and $0.3I$, where $I =$ In-sample length $/S_2$ and S_2 is the length of the longest cycle in the series (for the peak and hourly electricity demand $I = 20$ and for the half-hourly demand $I = 37$). For noise addition, levels $\vec{\sigma_1} = 0.1\vec{\sigma_b}$, $\vec{\sigma_2} = 0.2\vec{\sigma_b}$

and $\vec{\sigma_3} = 0.3\vec{\sigma_b}$ were used, and $\vec{\sigma_b}$ is the standard deviation of the bootstrapped original series. A normally distributed noise, $\vec{N}(0, \vec{\sigma_i})$, $i = 1, 2, 3$, was generated and added to the series, thus implying three different levels of uncertainty.

Performance of forecast combinations was assessed based on the original series. Comparison of our proposals are made against the following benchmarks: the seasonal naïve, a model in which the forecast for time period t and lead time h is $\hat{y}_{t+h|t} = y_{t+h-m}$, where m is the longest seasonal cycle; the average of point forecasts of all models in each pool; the base best model (denoted as *Base Seasonal* or *Base Dd. Seasonal*), which is an instance of the base model fitted with the original series (without noise addition or block swapping). The model selected has the lowest in sample RMSE within 100 trials obtained using different random starting points. For the single-seasonal time series the *ets* function from R *forecast* package (Hyndman, 2018) was used to fit an exponential smoothing benchmark (which will be called ETS). For the double-seasonal time series the functions *dshw* (an implementation of models proposed by Taylor, 2003) and *tbats* (De Livera et al., 2011) from the same package were used to fit double-seasonal exponential smoothing models. They are called Dshw and Tbats, respectively. Finally, the average between the base model and the best performing among ETS, Dshw and Tbats was calculated. It is denoted as Avg(Ets, Base Seasonal) for the first study, and analogously for the other two. Table 2 summarises the forecasting methods. The implementation column indicates the methods, equations or sources that were used.

Performance is assessed by using the Mean Squared Error $MSE = (1/n)\sum_i^n (y_i - \hat{y}_i)^2$, Symmetric Mean Absolute Percentage Error $SMAPE = (100/n)\sum_i^n \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$ and Geometric Mean Relative Absolute Error $GMRAE = \sqrt[n]{\prod_i^n \frac{|y_i - \hat{y}_i|}{|y_i - \hat{y'}_i|}}$, where n is the number of out-of-sample observations and $\hat{y'}_i$ is a benchmark forecast (Base Seasonal or Base Db. Seasonal). As results

Table 2: Forecasting methods

Applications	Forecasting method	Implementation
D, H, HH	Cluster based with noise addition* and block swapping**	Combination models in Section 2.3.1, (Equations 3, 4 and 5), with Equation 1 (single-seasonal) or 2 (double-seasonal) as base model.
D, H, HH	Genetic algorithms based with noise addition* and block swapping**	Combination models in Section 2.3.2, using Matlab® 2010 *ga* routine, with Equation 1 (single-seasonal) or 2 (double-seasonal) as base model.
D, H, HH	Seasonal naive	Forecast for period t and lead time h is $\hat{y}_{t+h\|t} = y_{t+h-m}$, where m is the longest seasonal cycle.
H, HH	Base double seasonal model	Model in Equation 2.
H, HH	Dshw	The function *dshw* from R *forecast* package (Hyndman, 2018) that implements double-seasonal models by Taylor (2003) was used.
H, HH	Tbats	The function *tbats* from R *forecast* package (Hyndman, 2018) that implements double-seasonal models was used.
H, HH	Average of Dshw and Base Db. Seasonal	
D	Base seasonal model	Model in Equation 1.
D	ETS (exponential smoothing benchmark)	The function *ets* from R *forecast* package (Hyndman, 2018) that implements exponential smoothing state space models was used.
D	Average of ETS and Base seasonal.	

D: Daily peak electricity demand, H: Hourly electricity demand, HH: Half-hourly electricity electricity demand.

* Structural combinations with three levels of noise addition and three levels of maximum number of clusters.

** Structural combinations with three levels of block swapping and three levels of maximum number of clusters.

for MSE and SMAPE are similar, we report only the later. To further assess the difference between the proposed approaches and the base models, a Wilcoxon test was perform to test the hypothesis $|e_{\text{model}}| < |e_{\text{base seasonal}}|$, where e_{model} is the out-of-sample error of a given model.

3.1. Model Confidence Set

Given the number of competing combinations, the Model Confidence Set, MCS, (Hansen et al., 2011) was estimated for each application. A MCS is a set of models that is constructed so that it will contain the best model with a given level of confidence $1 - \alpha$. We used $\alpha = 0.1$ in an R implementation by Catania (2014). We report the best models according to MCS. It is noteworthy that due to memory constraints of the R routines, the MCS was calculated with the full result set only in the first application. For the second application, 30% of the error set could be used and for the last application, 10%. The samples, however, were large, and all forecast horizons were included.

4. Applications

4.1. Forecasting daily peak electricity demand with structural combinations of Holt-Winters Forecasts

Structural combinations of Holt-Winters models are illustrated using data from Sunday 5 May 1996 to Saturday 30 November 1996 extracted from (Taylor et al., 2006). The first 20 weeks (140 observations) were used for training (fitting) and the remaining 10 weeks (70 observations) were used for evaluating the accuracy of forecasts up to 7 days ahead. The time series is depicted in Figure 2, where seasonality and time-varying volatility can be observed.

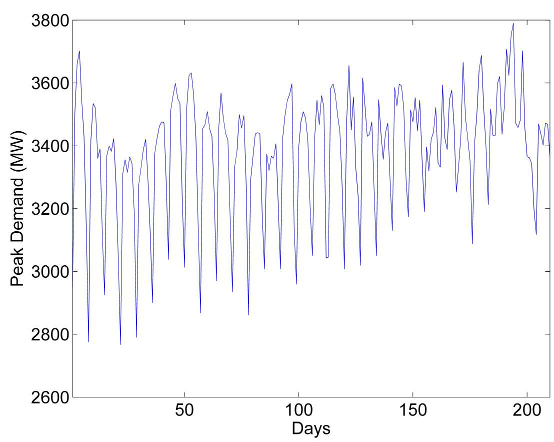

Figure 2: Daily peak electricity demand in Rio de Janeiro from Sunday, 5 May 1996 to Saturday, 30 November 1996.

4.1.1. Results

CB and GA combinations using both noise addition and block swapping at all 3 levels, the average forecast for each method-level combination and the benchmarks described in section 3 are considered. Figure 3 illustrates the performance of the best forecasting approaches and benchmarks. Further detail is provided in Table 3: the test on differences with respect to the base model; the rank within the Model Confidence Set, if a forecasting approach was included in it; the overall SMAPE and GMRAE with ranks and performance measurements for one horizon.

In general, improvements over the base best model, the average and the ETS benchmark suggest that multiplicative Holt-Winters can be combined with the proposed approach, and the combination is competitive against established benchmarks. For brevity, GA combinations are not reported, but perform relatively well and notably better when model diversity was induced through block swapping. By contrast, CB combinations do better with noise addition, but perform worse than the benchmarks in several forecast horizons. However, the combination CB(Noise, 2, 4), with a middle level of noise addition and a maximum of 4 clusters outperforms all benchmarks: with respect to base model, improvement in SMAPE ranges from 4.41% to 6.74% and between 7.39% and 8.54% in MSE. Concerning the forecast error distributions, forecast errors from all selected models pass the Shapiro–Wilk test of normality in all horizons (at 5% significance level).

Table 3: Best performers (out of 46) for daily peak load forecasting

Model	Diff.	MCS rank	SMAPE (rank)	GMRAE (rank)	SMAPE(GMRAE) $t+1$
CB(Noise,2,4)	7	1	2.927% (1)	0.895 (1)	2.394% (0.975)
CB(Swap,2,2)	7	3	3.031% (2)	0.903 (2)	2.503% (1.020)
GA(Swap,1,2)	7	2	2.779% (15)	1.000 (14)	2.502% (1.051)
Base Seasonal	NA		3.105% (7)	1.000 (21)	2.520% (1.000)
ETS	5		3.785% (45)	1.095 (41)	3.572% (1.570)
Avg(ETS, Base Seasonal)	7		3.307% (37)	0.980 (11)	2.906% (1.272)

Diff: number of horizons for which $|e_{\text{model}}| < |e_{\text{base seasonal}}|$ cannot be rejected at 5% significance.

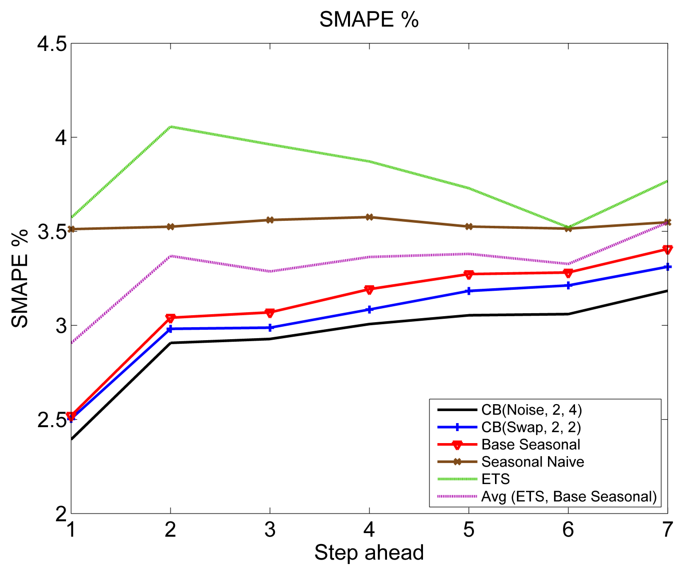

Figure 3: Best forecasting performance for daily peak electricity demand.

4.2. Forecasting hourly electricity demand with structural combinations of Double Seasonal Exponential Smoothing Forecasts

The time series of hourly electricity demand (Figure 4), which was used by Taylor et al. (2006) to assess performance of various univariate forecasting methods is considered. The first 20 weeks of data (equivalent to 3360 hourly observations) were used for training (fitting), and the remaining 10 weeks (equivalent to 1680 observations) were used for evaluating the accuracy of forecasts up to 24 hours ahead. Both the Tbats and Dshw benchmarks are considered.

4.2.1. Results

For this time series, structural combinations based on replicas of the time series generated via noise addition were poor. Their performance was on average at least 8% in SMAPE worse than the best base model. Hence, this section concentrates on the results obtained with block swapping, as illustrated in Figure 5, depicting SMAPE for all horizons.

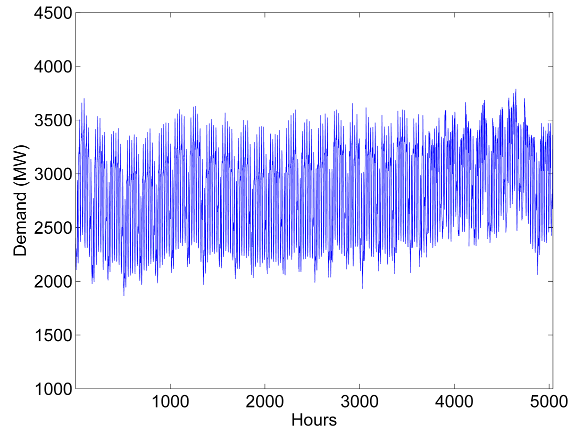

Figure 4: Hourly electricity demand in Rio de Janeiro for Sunday, 5 May 1996 to Saturday, 30 November 1996. source:Taylor et al. (2006)

Gains are very small with respect to the base model and the averages [2]. Both Dshw and Tbasts benchmarks are outperformed by the set of selected CB and GA combinations and do not appear in the MCS (Table 4). When the higher level of block swapping is considered, CB with 2 clusters and GA with 4 outperform the simple average, and thus are alternatives to the standard ensemble forecast.

Examining the forecast error distributions in each horizon, normality was rejected by the Shapiro-Wilk test. Out-of-sample errors for all models and benchmarks behave similarly, except those from Tbats. It was observed that the distributions change their central location for horizons 2 and 3, whereas for the remaining horizons, they are centred around zero. Thus, two and three step-ahead forecasts seem to be biased, and this could impact the performance of combinations.

[2] GA(Swap, 3, 8) reduces SMAPE by up to 0.34% compared to Avg(Swap, 3) and by up to 2.79% compared to the base model.

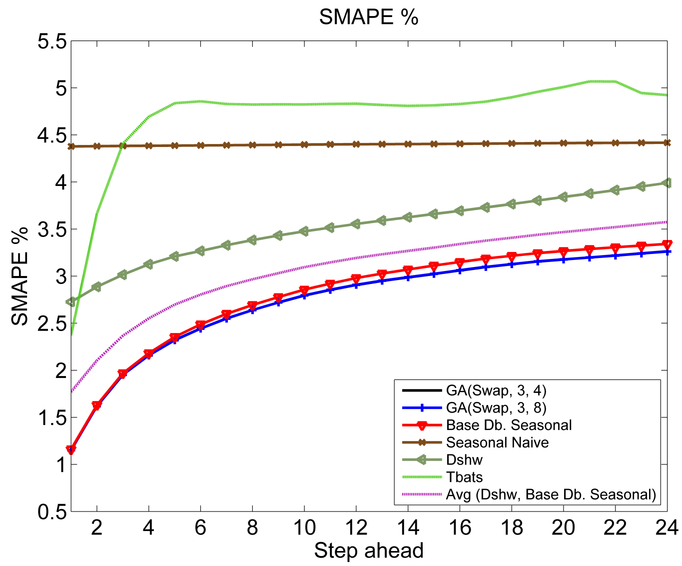

Figure 5: Best forecasting performance for hourly electricity demand.

Table 4: Best performers (out of 47) for hourly electricity demand.

Model	Diff.	MCS rank	SMAPE (rank)	GMRAE (rank)	SMAPE(GMRAE) $t+1$
CB(Swap,2,2)	24	3	2.725% (1)	0.991 (6)	1.178% (0.996)
GA(Swap,3,4)	24	2	2.732% (2)	0.984 (1)	1.153% (0.983)
GA(Swap,3,8)	24	1	2.732% (3)	0.984 (3)	1.151% (0.986)
Base Db. Seasonal	NA		2.796% (18)	1.000 (15)	1.160% (1.000)
Dshw	0		3.514% (38)	1.338 (40)	2.726% (2.483)
Tbats	0		4.698% (46)	1.679 (45)	2.368% (1.621)
Avg(Dshw, Base Db. Seasonal)	0		3.065% (25)	1.138 (25)	1.771% (1.583)

Diff: number of horizons for which $|e_{\text{model}}| < |e_{\text{base seasonal}}|$ cannot be rejected at 5% significance.

4.3. Forecasting half-hourly electricity demand with structural combinations of Double Seasonal Exponential Smoothing Forecasts

Half-hourly observations of electricity demand in England and Wales for the year 2016 (from Friday, 1 January 2016, to Saturday, 31 December 2016) were split into a training period consisting of 35 weeks and an evaluation period of 17.3 weeks (121 days). Forecasts up to 24 hours ahead are considered. [3] The time series comprising 17568 observations is depicted in Figure 6.

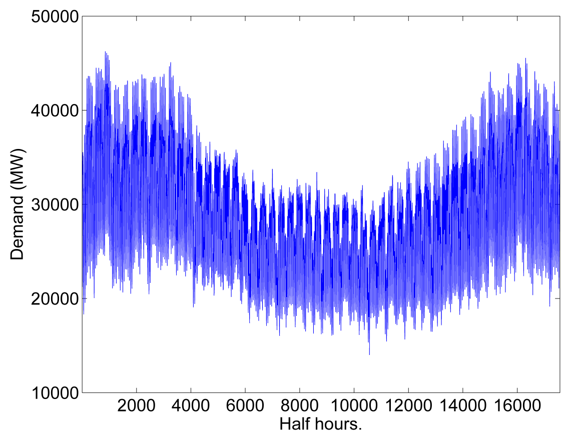

Figure 6: Half-hourly electricity demand in England and Wales, 1 January 2016 to 31 December 2016. source: National Grid

[3] Adjustments were made on 27 March and 30 October, when the clock went forward and backward. On the first date, when the clock went forward one hour, the resulting missing data points were linearly interpolated. On the second date, data for the repeated observations were averaged. Two additional missing points were linearly interpolated. Observations corresponding to public holidays (according to the Bank of England) and Christmas were also smoothed by replacing demand on each special day by the mean of the demand in the corresponding periods of the two adjacent weeks.

4.3.1. Results

When comparing the two approaches for diversity generation in ensembles, there are benefits from using noise addition over block swapping, specially for longer forecasting horizons (see Figure 7). Table 5 summarises results. The structural combinations outperformed both double seasonal benchmarks and the averages: GA(Swap,1,2) reduces SMAPE by up to 49% with restpect to Dshw and by up to 14% with respect to Avg(Swap, 1).

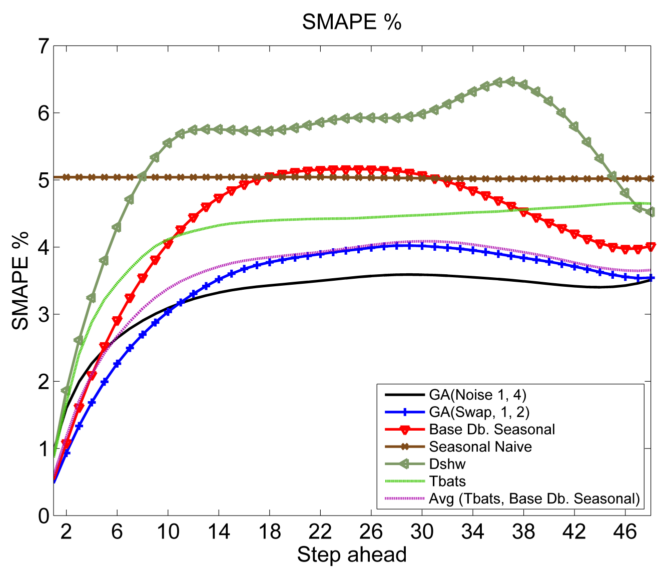

Figure 7: Best forecasting performance for England and Wales electricity demand.

Considering the addition of noise, the $GA(Noise, 1, 2)$, $GA(Noise, 1, 4)$ and $GA(Noise, 3, 2)$ outperformed the other benchmarks in most horizons. Given the middle level of noise, $CB(Noise, 2, 4)$ and $CB(Noise, 2, 8)$ perform well, but overall, GA combinations perform better. The block swapping approach worked better with GA combinations for the first forecast horizons, outperforming all benchmarks. They were consistently superior to the base model as illustrated by the tests on differences in Table 5.

Forecast error distributions rejected normality as per the Shapiro–Wilk

Table 5: Best performers (out of 47) for half-hourly electricity demand.

Model	Diff.	MCS rank	SMAPE (rank)	GMRAE (rank)	SMAPE(GMRAE) $t+1$
GA(Noise,1,4)	44		3.237% (1)	0.777 (2)	0.975% (1.828)
GA(Swap,1,2)	48	1	3.402% (13)	0.783 (5)	0.484% (0.889)
GA(Boot,3,4)	48	2	3.356% (8)	0.774 (1)	0.485% (0.892)
Base Db. Seasonal	NA		4.265% (45)	1.000 (44)	0.537% (1.000)
Dshw	0		5.383% (47)	1.262 (46)	0.899% (1.662)
Tbats	29		4.171% (44)	1.005 (45)	0.863% (1.627)
Avg(Tbats, Base Db. Seasonal)	45		3.546% (25)	0.852 (24)	0.598% (1.128)

Diff: number of horizons for which $|e_{\text{model}}| < |e_{\text{base seasonal}}|$ cannot be rejected at 5% significance.

test. Some of the ensemble combinations forecast error distributions that are better behaved than those of the best base model, thus suggesting that combinations from the ensembles can provide more robust forecasts.

5. Summary and conclusion

Ensembles have been applied to neural networks and exponential smoothing methods, thus exploring the general idea of creating diverse models under different conditions. This approach was applied here to the Holt-Winters and Multiplicative Holt-Winters-Taylor models. In contrast to neural networks, optimal parameters (structural descriptors) for these types of models tend to be homogeneous. Therefore, diversity was promoted by adding noise or swapping blocks of data to generate replicas of the time series. The latter approach is akin to the bootstrap performed by Bergmeir et al. (2016), but in this study the focus is on the time series rather than its components.

Three applications illustrated the performance of proposed combinations. Comparisons were made against the average of point forecasts in the ensembles, the base best model (building block in each ensemble) and other suitable benchmarks. Results are encouraging, but also highlight the need for further investigation, as several questions also emerged from the study.

These relate to the choice of variation induction mechanisms, and how to optimise the number of clusters and other factors in the procedures. We also note that additional analysis with simulated data suggested that the length of the seasonal cycles may influence performance and thus the choice of structural combination.

For the single-seasonal daily time series series (peak electricity demand), improvements over both the average and the best model in the pool were observed, but the strategy of noise addition for model diversity worked better than block swapping in producing competitive structural combinations. Results suggest that CB combinations are better at exploiting model variations coming from noise addition in order to improve performance on this series. GA combinations, on the other hand, seem to perform well under both approaches. Additionally, CB combinations are volatile while GA are not. These observations deserve future investigation, as different clustering procedures and measures of memberships to clusters can also be considered.

In the case of the first double-seasonal time series (hourly electricity demand in Rio de Janeiro), improvement over the average forecast in the ensembles produced with noise addition was more common in CB combinations than in GA. By contrast, for the second time series (half-hourly electricity demand in England and Wales), improvement over the average was easier for GA combinations than for CB under both noise addition and block swapping. For some forecast horizons, 14% reduction in SMAPE were observed for several horizons, with significant gains by both types of structural combinations, independently of how replicas of the original time series were generated.

Overall, the results suggest that structural combinations can outperform standard ensemble forecasts, and underscore how robust exponential smoothing models can be. Given the larger number of observations in the last application, which led to better results, further investigation should consider how sensitive are these combinations to the number of observations and cycles in the in-sample period. With the availability of larger data sets and increasing computing power, the methodology proposed can be also applied to different forecasting problems with seasonal data (e.g. hourly road traffic, demand for public transportation, access to websites, volumes in call centres). Finally, these structural combinations can be extended to include other types of models and support the development of hybrid ensembles.

The results obtained can be viewed from the perspective of a learning process, interpreted as a link between a problem space and a solution space (Kasabov, 1996, p. 332). For different problems (data) there are different mappings (forecasting algorithms) that lead to a solution. In this research the fitting of a forecasting model was used instead of a learning algorithm (as in neural networks), and variation was introduced into the problem by altering the data. These variations led to different mappings. Subsequently, such mappings were combined, structurally. When using cluster-based combinations (which showed volatility), the forecasting performance could undergo favourable jumps when noise was added to the data and therefore improve markedly over benchmarks. This could be interpreted as a jump in the search of a problem-solution mapping. When data were diversified through block swapping, the mappings (or fitted models) provided a more stable performance, closer to the average and the base model. The first situation is specially observable in CB combination, when applied to the single-seasonal time series. The second situation is observed in various CB and GA combinations in both hourly and half-hourly electricity demand time series.

The selection of structural combinations, as proposed in this study, can be made by using a rule of thumb: if significant forecast improvement by a cluster based combination with respect to the base best model and other benchmarks is observed in the first horizons, then the cluster-based combination is preferred. If insignificant improvement is obtained for the first horizons, either the base best model, a well-performing benchmark or a GA combination is preferred.

Future research may also consider state-space models, due to their greater use of model structural information. Subsequent studies can also explore the automation and comparison of alternative design decisions, such as the maximum number of clusters and the number of models selected per cluster.

References

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners* volume 30. Springer.

Arora, S. (2013). *Time Series Forecasting with Applications in Macroeconomics and Energy*. Ph.D. thesis Somerville College, University of Oxford.

Bakker, B., & Heskes, T. (2003). Clustering ensembles of neural network models. *Neural networks : the official journal of the International Neural Network Society*, *16*, 261–269.

Barak, S., & Sadegh, S. S. (2016). Forecasting energy consumption using ensemble arima–anfis hybrid algorithm. *International Journal of Electrical Power & Energy Systems*, *82*, 92 – 104.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*, 451–468.

Bergmeir, C., Hyndman, R., & Benítez, J. (2016). Bagging exponential smoothing methods using stl decomposition and box–cox transformation. *International Journal of Forecasting*, *32*, 303 – 312.

Brown, W. M., Gedeon, T. D., & Groves, D. I. (2003). Use of noise to augment training data: a neural network method of mineral–potential mapping in regions of limited known deposit examples. *Natural Resources Research*, *12*, 141–152.

Catania, L. (2014). *MCSprocedure*. URL: `https://www.rdocumentation.org/packages/MCS/versions/0.1.3/topics/MCSprocedure` r package version 0.1.3.

Chan, C. K., Kingsman, B. G., & Wong, H. (2004). Determining when to update the weights in combined forecasts for product demand—an application of the cusum technique. *European Journal of Operational Research*, *153*, 757 – 768. EURO Young Scientists.

Che, J. (2015). Optimal sub-models selection algorithm for combination forecasting model. *Neurocomputing*, *151*, 364–375.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559 – 583.

Clements, M. P., & Harvey, D. I. (2010). Forecast encompassing tests and probability forecasts. *Journal of Applied Econometrics*, *25*, 1028–1062.

Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting*, *27*, 635–660.

De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, *106*, 1513–1527.

De Menezes, L., & Bunn, D. (1998). The persistence of specification problems in the distribution of combined forecast errors. *International Journal of Forecasting*, *14*, 415 – 426.

De Menezes, L., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*, 190–204.

Deutsch, M., Granger, C. W., & Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting*, *10*, 47–57.

Donaldson, R. G., & Kamstra, M. (1996). Forecast combining with neural networks. *Journal of Forecasting*, *15*, 49–61.

Fiordaliso, a. (1998). A nonlinear forecasts combination method based on Takagi-Sugeno fuzzy systems. *International Journal of Forecasting*, *14*, 367–379.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*, 1–67.

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert*, *6*, 46–51.

Goh, A. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, *9*, 143 – 151.

Granger, C. (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press.

Hahn, H., Meyer-Nieberg, S., & Pickl, S. (2009). Electric load forecasting methods: Tools for decision making. *European Journal of Operational Research*, *199*, 902 – 907.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*, 453–497.

Hippert, H., Pedreira, C., & Souza, R. (2001). Neural networks for short-term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems*, *16*, 44–55.

Hyndman, R., Koehler, A., Ord, J., & Snyder, R. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Series in Statistics. Springer Berlin Heidelberg.

Hyndman, R. J. (2018). *forecast: Forecasting functions for time series and linear models*. URL: `http://github.com/robjhyndman/forecast` r package version 8.3.

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*. OTexts.

Jang, J., Sun, C., & Mizutani, E. (1997). *Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence*. Prentice Hall.

Jing, B.-Y. (1997). On the relative performance of the block bootstrap for dependent data. *Communications in Statistics-Theory and Methods*, *26*, 1313–1328.

Kasabov, N. (1996). *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*. A Bradford book. MIT Press.

Khwaja, A., Zhang, X., Anpalagan, A., & Venkatesh, B. (2017). Boosted neural networks for improved short-term electric load forecasting. *Electric Power Systems Research*, *143*, 431 – 437.

Kolassa, S. (2011). Combining exponential smoothing forecasts using akaike weights. *International Journal of Forecasting*, *27*, 238 – 251.

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*, 291 – 302.

Li, S., Goel, L., & Wang, P. (2016). An ensemble approach for short-term load forecasting by extreme learning machine. *Applied Energy*, *170*, 22 – 29.

Matijaš, M., a.K. Suykens, J., & Krajcar, S. (2013). Load forecasting using a multivariate meta-learning system. *Expert Systems with Applications*, *40*, 1–11.

Nowotarski, J., Liu, B., Weron, R., & Hong, T. (2016). Improving short term load forecast accuracy via combining sister forecasts. *Energy*, *98*, 40–49.

Ord, K., Fildes, R. A., & Kourentzes, N. (2017). *Principles of Business Forecasting*. (1st ed.). Wessex Press Publishing Co.

Pai, P.-F., & Hong, W.-C. (2005). Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. *Electric Power Systems Research*, *74*, 417 – 425.

Pesaran, M. H., & Pick, A. (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics*, *29*, 307–318.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, *268*, 545 – 554.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430.

Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, *54*, 799–805.

Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, *204*, 139–152.

Taylor, J. W., & Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, *19*, 57–70.

Taylor, J. W., & Majithia, S. (2000). Using combined forecasts with changing weights for electricity demand profiling. *The Journal of the Operational Research Society*, *51*, pp. 72–82.

Taylor, J. W., de Menezes, L. M., & McSharry, P. E. (2006). A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, *22*, 1–16.

Timmermann, A. (2006). Chapter 4 forecast combinations. (pp. 135 – 196). Elsevier volume 1 of *Handbook of Economic Forecasting*.

Xiong, L., Shamseldin, A. Y., & O'Connor, K. M. (2001). A non-linear combination of the forecasts of rainfall-runoff models by the first-order takagi–sugeno fuzzy system. *Journal of Hydrology*, *245*, 196 – 217.

Zhang, G. P. (2007). A neural network ensemble method with jittered training data for time series forecasting. *Information Sciences*, *177*, 5329–5346.

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, *137*, 239–263.